Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

# Speaker-basis Accent Clustering Using Invariant Structure Analysis and the Speech Accent Archive

*N. Minematsu*[†], *S. Kasahara*[†], *T. Makino*[‡], *D. Saito*[†], *K. Hirose*[†]

† The University of Tokyo, Tokyo, Japan
‡ Chuo University, Tokyo, Japan
{mine,kasahara,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp, mackinaw@tamacc.chuo-u.ac.jp

## Abstract

English is the only language available for global communication and is used by 1.5 billions of speakers. It is also known to have a large diversity of pronunciation due to the influence of speakers' mother tongue, called accents. Our project aims at creating a global and speaker-basis map of English accents to be used in learning World Englishes as well as research studies of World Englishes [1, 2]. Creating the map, i.e., speaker-basis accent clustering, mathematically requires a distance matrix in terms of accents among all the speakers considered, and technically requires a method of predicting the accent distance between any pair of the speakers by using their speech samples only. In [3, 4], our first trials were presented, where invariant structure analysis was effectively used for feature extraction. However, some technical problems were found through the experiments and in this paper, recent progresses are presented with additional explanation on the invariant structure, which were omitted in [3, 4] due to space limitations. Use of the invariant structure and Support Vector Regression shows a striking performance of distance prediction in a speaker-pair-open mode but the performance is not sufficient in a speaker-open mode.

## 1. Introduction

In English classes, British or American English (BE or AE) is often presented as a reference, which learners try to imitate. It is widely accepted, however, that native-like pronunciation is not always needed for smooth communication. Due to the influence of the learners' mother tongue, those from different regions inevitably have different accents in their pronunciation of English. Recently, more and more teachers accept the concept of World Englishes (WE) [1, 2] and they regard BE and AE just as two major examples of accented English. Diversity of WE can be found in various aspects such as dialogue, syntax, pragmatics, lexical choice, spelling, pronunciation, etc. Among these kinds of diversity, this paper focuses on pronunciation. If one takes the concept of WE as it is, he can claim that there does not exist the standard pronunciation of English. If it has to be defined statistically, Chinese English may be the standard pronunciation of English because of the large number of its users [5]. The authors expect that there will be a great interest in how one type of pronunciation compares to other varieties, not in how that pronunciation is incorrect compared to the standard pronunciation.

The ultimate goal of our project is creating a global map of WE on a speaker basis for each of the speakers to know how his pronunciation is located in the diversity of WE. If the speaker is a learner, he can then find easier-to-communicate English conversation partners, who are supposed to have a similar kind of pronunciation. If he is too distant from many of other varieties, however, he may have to correct his pronunciation for the first time to achieve smoother communication with these others.

In this paper, we use the Speech Accent Archive (SAA) [6], which provides speech samples of a common elicitation paragraph read by more than 18 hundred speakers from all over the world. The SAA also provides IPA-based narrow transcripts of all the samples, which can be used for training an accent distance predictor. To calculate the accent distance between two speakers of the SAA, [7, 8] proposed a method of comparing two IPA transcripts using the edit distance. Although it was shown that the calculated distances had reasonable correlation with the accent distances perceived by human listeners, unlabeled data, i.e., raw speech, were not handled in [7, 8]. Recently, we proposed a method of predicting the accent distance only using spoken paragraphs of the SAA [3, 4].

The technical challenge is how to make the prediction independent of irrelevant but inevitably involved factors such as differences in age, gender, channel, background noise, etc. In studies of speech recognition and language (accent) identification, speaker identity is often treated as a hidden or latent variable and, through collecting samples over a large number of speakers, speaker identity becomes unseen in distributions of the variables of interest[1]. In our study, however, this strategy is not adequate because we're attempting *speaker-basis* accent clustering, where the unit of accent is a speaker and accent modeling has to be done for a speaker, not for a speaker cluster. To solve this extremely challenging problem, we suppress speaker identity in acoustic observations of speech as phase and pitch harmonics can be effectively removed from them. For this aim, in [3, 4], we used invariant pronunciation structure analysis [9, 10, 11, 12] for feature extraction and support vector regression (SVR) for prediction. In training the predictor, reference distances had to be prepared. In [3, 4], IPA-based phonetic distances calculated through dynamic time warping (DTW) of two IPA transcripts were used. In [3, 4], however, all the experiments were carried out in a speaker-pair-open mode and in this paper, new results in a speaker-open mode will be presented.

## 2. The speech accent archive

This corpus is composed of read speech samples of more than 18 hundred speakers and their corresponding IPA narrow tran-

[1] The term of "speaker-independent" is often used to indicate *statistical* independence. Since the theory of probability defines $P(a) = \sum_b P(a, b)$, through collecting samples, any variable can be treated as hidden or latent variable. Speaker-independent HMMs are trained by this strategy using a large speech corpus to make speaker identity unseen in the distribution. In this paper, we focus on another kind of independence, which should be referred to as *physical* independence, where a variable can be suppressed or separated from acoustic observations.

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

[pli:z̥ kɔl ə̆stɛl:ʌ as hɛr tu brɪŋ diz θɪŋs wɪθ hɛr frʌm ðə stɑɹ sɪks spu:nz ʌv̥ fɹɛʃ ə̆sno pi:z faɪv̥ θɪk ə̆slɛb̥s ʌv blu: ʧi:z æn meɪbi: eɪ snæk˺ foɹ hɛɹ bɹʌðɜ bɑb˺ wĭ ɑlso nid˺ eɪ smɑlˠ plæstɪk˺ ə̆s̬neɪk æn eɪ bɪg̊ tʰɔɪ fɹɔg˺ fɔɹ ðə kɪdz̥ ʃi kɛn ə̆sku:b˺ ði:z θɪŋs ɪntu θri: ɹɛd˺ bægs æn ə wĭl go: mitʰ hɛɹ wɛnzdeɪ æd˺ d̪ə tɹeɪn ə̆steɪʃən]

Figure 1: The SAA paragraph and an example of transcription

scripts. The speakers are from all over the world and they read the common elicitation paragraph, shown in Figure 1, where an example of IPA transcription is also presented. The paragraph contains 69 words and can be divided into 221 phoneme instances using the CMU dictionary as reference [13]. The IPA transcripts are used to prepare reference inter-speaker accent distances, which will be adopted as target of prediction using SVR in our study. This is because IPA transcription is done through phoneticians' ignorance of non-linguistic and acoustic variations involved in utterances such as differences in age, gender, channel, etc. It should be noted that the recording condition in the corpus varies from sample to sample because data collection was done voluntarily by those who had interest in joining the SAA project. To create a suitable map automatically, these non-linguistic variations have to be cancelled without collecting samples covering those variations well. This is because, in this work, each accent has only a single sample.

Use of read speech, not spontaneous speech, is considered to reduce accent diversity because read speech may show only controlled or artificial diversity. In [14], however, English sentences read by 200 Japanese university students still showed a very large diversity. Further in [15], a listening test of English sentences read by Japanese was done by using a large number of American listeners as subjects. The results showed that the intelligibility of the read sentences covered a very wide range. Following these facts, we considered that read samples can still show well how diverse World Englishes are in terms of accent.

It is well-known that accent diversity is found in both of the segmental and prosodic aspects. In this paper, however, we will prepare reference accent distances by using IPA transcripts, meaning that prosodic diversity will be lost. We do not claim that the prosodic diversity is minor but, as was shown in [16], clustering only based on the segmental aspect seems able to show adequately how diverse World Englishes are in terms of accent. Reference distances defined with both segmental and prosodic features will be addressed in a future work.

In this study, only the data with no word-level insertion or deletion were used. The speech files that contained exactly 69 words were automatically selected. Some of them were found to include a very high level of background noise, and we manually removed them. Finally, 370 speakers' data were used and the number of speaker pairs is 68,265 (=$_{370}C_2$=370×369 / 2).

## 3. Invariant speech structure

### 3.1. Infants' sensitivity to the sound system of a language

How to suppress speaker identity in acoustic observations? Phase characteristics are often removed in speech analysis because human hearing is rather insensitive to phase differences. Pitch harmonics are also removed in many applications because phoneme identity is independent of pitch unless tonal languages are dealt with. Accents of English pronunciation are independent of speakers' age and gender but their acoustic observations are inevitably altered due to these factors.

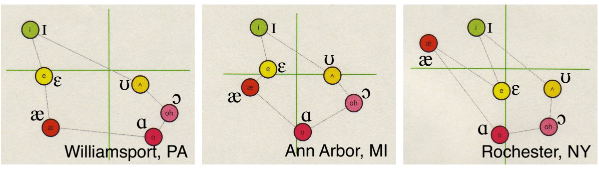

Figure 2: Dialect-specific vowel distributions in AE [21]

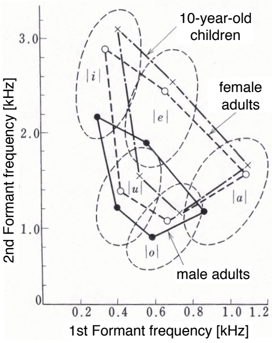

Figure 3: Japanese vowel distribution patterns [22]

What is the speech pattern that is *physically* independent of these factors? Our answer to this question is the invariant structure in speech, which was proposed in [9, 10, 11]. This technical proposal was inspired by recent studies of human infants [17, 18] and animals [19] and by classical studies of structural phonology [20]. When we consider infants' performance of vocal imitation, we can say that they are very insensitive to speaker identity when acquiring a language. When infants imitate their parents' utterances, they do not imitate them acoustically. Their vocal imitation is not impersonation. It seems that they imitate the *physically* speaker-independent pattern embedded in the parents' utterances. It is interesting that researchers of animal sciences describe that vocal imitation is rarely found in animals and only birds, dolphins, and whales imitate vocally. But their imitation is basically acoustic imitation [19].

In [17, 18], it was experimentally shown that infants are very sensitive to distributional properties of speech sounds exposed to them. This sensitivity is easy to understand when we see dialectal differences found in a language. Figure 2 shows several distribution patterns of six vowels of AE dialects [21]. The vowels are plotted on F1/F2 planes after vocal tract length normalization. It is well-known that different dialects show different vowel distribution patterns. When an infant is born and brought up in an geographical area, it inevitably acquires the regional accent (distributional pattern) of that area.

Figure 3 shows Japanese vowel distributions of male adults, female adults, and 10-year-old children [22]. As we described above, infants do not impersonate their parents or caretakers but they do learn the sound distribution pattern. Considering this performance, we can say that infants are not sensitive to absolute properties of speech sounds but sensitive to relational or distributional properties in them. Putting it in another way, infants are sensitive to the sound system of a language [20], which seems to be physically speaker-independent.

### 3.2. Derivation of invariant speech structure

What is the simplest definition of a (sound) system? Geometrically speaking, the shape of a three-point structure (a triangle) can be defined simply by the length of the three edges. What

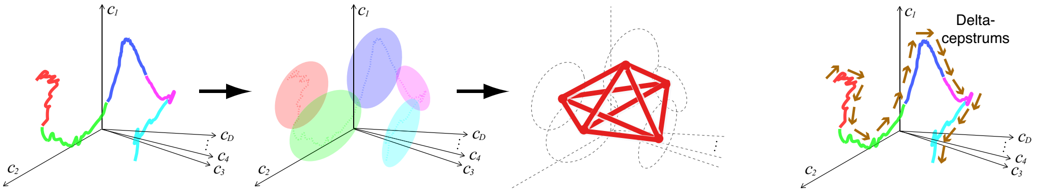

Figure 5: The speech structure composed only of $f$-divergences

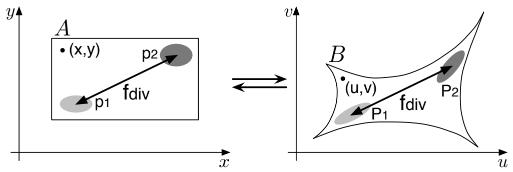

Figure 4: Invariance of $f$-divergence against topological deformation of manifolds (shapes)

about an $n$-point structure? In this case, the length of all the edges including the diagonal edges, i.e., the distance matrix of the $n$ points, can define the shape of that structure. If the matrix calculated from the sounds generated by a speaker and that of the same kinds of sounds from another speaker is the same, we can say that those matrices are speaker-invariant or *physically* speaker-independent and that infants seem to be sensitive to the matrix properties. How can one measure the length of an edge of an $n$-point structure in a speaker-invariant way?

In [9, 10, 11], we answered this question mathematically. Speaker difference is modeled as space mapping in studies of voice conversion. In [10], we proved that $f$-divergence between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency) and that any invariant metric of the integral form has to be written in the form of $f$-divergence (necessity). $f_{\text{div}}$ is a distance metric between two distributions on measurable space $\mathcal{X}$, $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$, as

$$f_{\text{div}}(p_1, p_2) = \int_{\mathcal{X}} p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}, \qquad (1)$$

where $g(t)$ is a convex function for $t > 0$. If we take $t\log(t)$ or $-\log(t)$ as $g(t)$, $f_{\text{div}}$ becomes KL-divergence. When $\sqrt{t}$ is used for $g(t)$, $-\log(f_{\text{div}})$ becomes Bhattacharyya distance (BD). Figure 4 shows two spaces (shapes) which are deformed into each other through an invertible and differentiable transform. An event is described not as a point but as a distribution. Two events of $p_1$ and $p_2$ in $A$ are transformed into $P_1$ and $P_2$ in $B$. Generally speaking, the two spaces are closed manifolds and the invariance of $f$-divegence is always satisfied [10].

$$f_{\text{div}}(p_1, p_2) \equiv f_{\text{div}}(P_1, P_2). \qquad (2)$$

In [9, 10, 11], we used the BD as one of the $f_{\text{div}}$ metrics. If an input utterance is represented as BD-based distance matrix by using only the distributions found in that utterance, then, the matrix is an invariant representation of that utterance. Figure 5 shows the process of deriving the invariant structure from an input utterance. The utterance in a feature space, such as cepstrum space, is a sequence of feature vectors and it is converted into a sequence of distributions through automatic and unsupervised segmentation. Here, any speech event is characterized as distribution. The BD is calculated from any pair of distributions and we can get an invariant distance matrix for that utterance. This matrix-based invariant representation is called speech structure [9, 10] or pronunciation structure [11, 12].

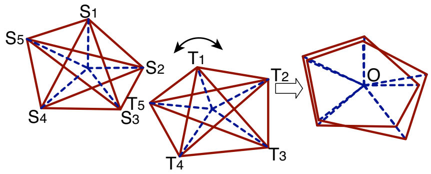

Figure 6: Structural comparison of two structures

Here, we should note that velocity vectors, relative and directional changes at each point on the time axis (See the right-hand side of Figure 5), are not good candidates for speaker-invariant features. This is because cepstrum normalization in terms of the vocal tract length can be approximated as multiplication of a rotational matrix by the original cepstrum vectors [23, 24] and therefore, the direction of velocity vectors are strongly dependent on the vocal tract length [24].

If the acoustic feature of interest is a one-dimensional feature, such as fundamental frequency, however, since rotation is geometrically impossible, velocity vectors can become invariant. Perception of relative and directional changes in pitch is often called relative pitch in musicology and, owing to this, one can perceive syllable names, not pitch names, of Do, Re, Mi,..., in a key-invariant way. We can say that $f_{\text{div}}$-based distance matrix is an extended concept of relative pitch, that will be relative *timbre*. Because pitch is one-dimensional and timbre is multi-dimensional, invariance can be found as local and directional and changes in the former but in the latter, it can found only as local contrasts or distant contrasts between acoustic events. We can say that our invariant structure is a general solution of finding invariance in dynamics of multi-dimensional features.

It is interesting that animals do not have relative pitch and therefore, a melody and its transposed version are just two different sound streams [25]. Considering animals' performance of vocal imitation and melody perception, it seems that their perception of sounds is very absolute and non-robust.

### 3.3. Structural comparison as computational shortcut

In [24], we showed that vocal tract length normalization can be approximated as rotating a cepstrum trajectory: $c'{=}Ac$, where $A$ is a rotational matrix. It is easy to describe that microphone normalization can be characterized in the cepstrum space as adding a constant vector $b$: $c'{=}c{+}b$. Considering this geometrical property, comparison of two $n$-point structures ($n{\times}n$ matrices) can be done by shifting and rotating a structure so that both the structures can be overlapped the most (See Figure 6). Difference between the two structures can be quantified as summation of distances between each corresponding pair of points between the two structures after overlapping, where rotation and shift mean vocal tract length normalization and microphone normalization, respectively. [26] shows experimentally that this difference is approximately proportional to the Euclidean distance between the two matrices, which is calculated by viewing a matrix as a vector. In other words, structural comparison can give

Table 1: Vowel substitution table

| Japanese vowels ↔ | English vowels |
|---|---|
| a | ɑ, ʌ, æ, ɚ, ə |
| i | i, ɪ |
| u | u, ʊ |
| e | ɛ |
| o | ɔ |

Table 2: 8 patterns of vowel substitution

| | ɑ | æ | ʌ | ə | ɚ | ɪ | i | ʊ | u | ɛ | ɔ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | J | J | J | J | J | J | J | J | J | J | J |
| P2 | A | A | A | A | A | J | J | J | J | J | J |
| P3 | J | J | J | J | J | A | A | A | A | A | A |
| P4 | A | A | J | J | J | A | A | J | J | A | A |
| P5 | J | J | A | A | A | J | J | A | A | J | J |
| P6 | A | J | A | J | A | J | J | J | J | A | A |
| P7 | J | A | J | A | J | A | A | A | A | J | J |
| P8 | A | A | A | A | A | A | A | A | A | A | A |

A : American English pronunciations are used.

J : Japanese vowels are substituted.

us acoustic distance between two utterances after normalization although no normalization was done explicitly. In this sense, structural comparison can be viewed as computational shortcut.

### 3.4. Use of speech structure to cluster simulated learners

In [27], we applied the pronunciation structure analysis to cluster simulated Japanese learners of English. We collected vowel samples of Japanese and AE from 12 Japanese returnees from America, who are referred to as speakers A to L. By using the vowel samples of a speaker, we built multiple vowel structures for that speaker. Here, by replacing some of the AE vowels by Japanese ones, we simulated different types of Japanese accented English vowels of the same speaker. Table 1 shows the substitution table between Japanese vowels and AE ones and Table 2 shows eight patterns used for vowel substitution. P8 uses the AE vowels uttered by returnee speakers and P1 substitutes Japanese vowels for all the AE ones, which simulates completely Japanese accented pronunciation of AE vowels. We had 12 speakers (A to L) and 8 substitution patterns (1 to 8), resulting in 96 different vowel systems in total.

By modeling each vowel segment as Gaussian distribution, an 11×11 matrix was built for each of the 96 vowel systems. By calculating distance between a vowel system and another, we can get a 96×96 distance matrix, which can cluster the 96 systems. In [27], the following two metrics were used.

$$D_1(S,T) = \sqrt{\tfrac{1}{11}\textstyle\sum_{i<j}(S_{ij}-T_{ij})^2} \qquad (3)$$

$$D_2(S,T) = \sqrt{\tfrac{1}{11}\textstyle\sum_{i} BD(v_i^S, v_i^T)} \qquad (4)$$

$X_{ij}$ is a $(i,j)$ element of speaker $X$'s matrix, which is calculated as square root of BD between vowels $i$ and $j$ of speaker $X$. $v_i^X$ is vowel $i$ of speaker $X$. $D_1(S,T)$ is the structural Euclidean distance between two speakers $S$ and $T$ by simply regarding their distance matrices as vectors. On the other hand, $D_2(S,T)$ corresponds to absolute difference between the corresponding vowels of speakers $S$ and $T$. Using $D_1$ and $D_2$, we can get two different 96×96 distance matrices.

Figure 7 and Figure 8 show two dendrograms generated from the two 96×96 distance matrices. The upper shows the result of using $D_1$ and the lower shows that of using $D_2$. It is clearly illustrated that the upper is accent clustering and the lower is speaker clustering. In the upper, although accept gap

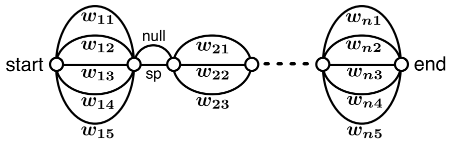

Figure 9: An example of word-based network grammar

was calculated in a unsupervised mode, speaker identity is suppressed or removed effectively by extracting the invariant structure from utterances of the speaker.

The invariant structure analysis was applied to quantify accent gap between a specific teacher and students in a supervised mode using ridge regression [12]. In [3, 4], the analysis was applied again to quantify accent gap among many native and non-native speakers toward creating a speaker-basis accent map of World Englishes, where SVR was used as regression model and IPA-based accent distance was used as reference.

## 4. Reference inter-speaker accent distance

To train an accent distance predictor, reference inter-speaker distances were needed, which were also used to evaluate the trained predictor. Following [7, 8], the reference distance between two speakers was calculated through DTW of their IPA transcripts. Since all the transcripts contained exactly the same number of words, word-level alignment was easy and we only had to treat phone-level insertions, deletions, and substitutions between a word and its counterpart.

Since DTW-based alignment of two IPA transcripts needs the distance matrix among all the existing IPA phones in the SAA, we prepared it in the following way. Since the number of kinds of IPA phones found in the narrow transcripts of the SAA was larger than 300, we focused on the most frequent 153 phones that can cover 95% of all the phone instances. Then, we asked an expert phonetician to pronounce each of the 153 phones twenty times. Using the recorded data, a speaker-dependent three-state HMM was built for each phone, where each state was modeled as Gaussian distribution. Then, for each phone pair, the phone-to-phone distance was defined as the average of three state-to-state Bhattacharyya distances. The other 5% of the phones were all with a diacritical mark. For each of them, we substituted the HMM of its base phone.

Using the distance matrix among all the kinds of phones in the SAA, word-based DTW was conducted to compare a word and its counterpart in IPA transcripts. The accumulated distance was normalized by the number of phones in the word pair and the normalized distances were summed for all the 69 words. This final distance was used as reference distance. Detailed configuration of our DTW, such as local paths and penalty scores, is found in [16] as well as a result of bottom-up clustering of a part of the SAA using IPA-based DTW.

## 5. Baseline systems

For comparison, we built two baseline systems, which corresponds directly to an automated version of the reference distance calculation procedure described in Section 4.

The calculation procedure is composed of two steps: 1) IPA manual transcription and 2) DTW alignment for distance calculation. The first process was replaced with automatic recognition of phonemes in input utterances[2]. Here, we used a phoneme

[2] As far as we know, there does not exist an automatic recognizer of IPA phones with a diacritical mark.

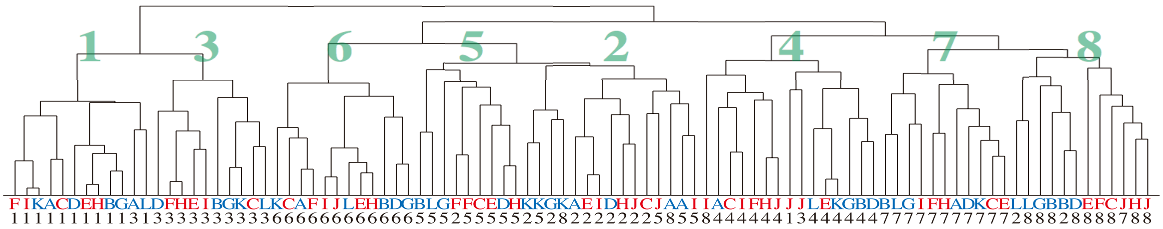

Figure 7: Clustering the 96 vowel structures based on structural comparison ($D_1$)

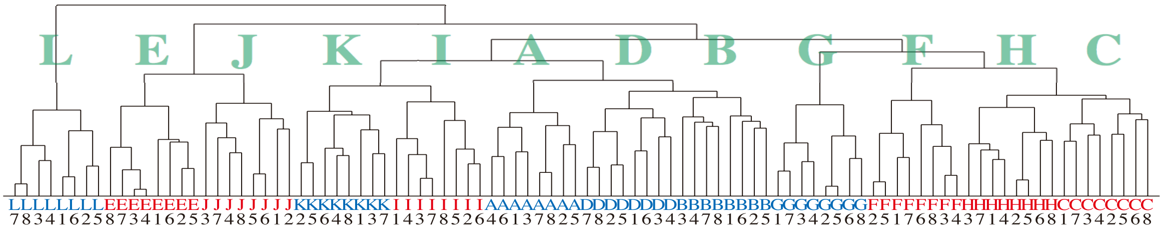

Figure 8: Clustering the 96 vowel structures based on absolute comparison ($D_2$)

recognizer of AE in this study. Using all the utterances of the 370 speakers as training data, monophone HMMs were trained by adopting the WSJ-based HMMs [28] as initial model. For this training, each IPA transcript was converted into its AE phoneme transcript. This conversion was done by preparing a phone-to-phoneme mapping table with special attention paid to conversion from two consecutive vowels to an AE diphthong.

Since IPA transcription is based on phones and the HMMs were trained based on phonemes, even if we could have a perfect phoneme recognizer, generated transcripts have to be phonemic versions of IPA transcripts. Phone to phoneme conversion is an abstraction process and some detailed phonetic information will be lost inevitably. Our first baseline system used IPA-based phonemic transcripts as output from an oracle system and the accent distance was calculated by comparing two phonemic transcripts based on DTW. Here, the phoneme-to-phoneme, not phone-to-phone, distance matrix was needed and prepared by using the WSJ-based HMMs. Our second system used a real phoneme recognizer with word-based network grammar that can cover all the pronunciation variations found in the 370 speakers. Figure 9 shows an example of the network grammar of an $n$-word sentence, where $w_{ij}$ denotes the $i$-th word spoken with the $j$-th pronunciation. A short pause can be inserted at word boundary. In the second system, the generated transcripts often included recognition errors.

The correlation between the IPA-based inter-speaker reference distances and the phoneme-based distances obtained from the first system was 0.829, meaning that information loss exists to some degree. The phoneme recognition accuracy of the second system was 73.5% but the correlation was so low as 0.458. This clearly indicates that recognition errors are very fatal.

## 6. Proposed method and experiments

### 6.1. Paragraph-based pronunciation structure

In this paper, the pronunciation structure was extracted from each of spoken paragraphs of the SAA. Here, the paragraph-based speaker-independent HMM was trained at first, which was used as Universal Background Model (UBM). Then, it was adapted through MAP adaptation to each speaker. The initial model for the UBM-HMM was prepared by concatenating AE phoneme HMMs trained from the WSJ corpus [28] by referring to the phoneme sequence derived from the SAA paragraph. The initial model was updated through ML-based parameter reestimation by using all the 370 available speakers of the SAA. This UBM-HMM was then adapted to each of the 370 speakers. Figure 10 schematizes the procedure adopted in this paper to calculate the pronunciation structure. The number of states of the paragraph-based HMM is $3N$, where $N$ is the number of phonemes of the SAA paragraph (=221). For each adapted paragraph-based HMM, the averaged BD between every pair of the phoneme instance HMMs was calculated, where the $i$-th phoneme instance HMM in the paragraph HMM is the three-state HMM spanning from the $(3i-2)$-th state to the $3i$-th state of that paragraph HMM. Finally in Figure 10, the pronunciation structure of a spoken paragraph of the SAA was obtained as 221×221 distance matrix. As illustrated in Figure 11, from two distance matrices of speakers $S$ and $T$, we derived a difference matrix $D$ to characterize the accent gap between them.

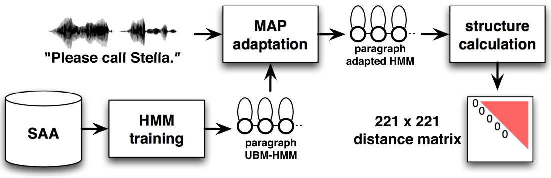

Figure 10: Procedure to calculate the pronunciation structure

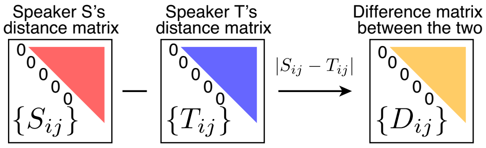
Speaker S's distance matrix
Speaker T's distance matrix
Difference matrix between the two
$\{S_{ij}\}$
$\{T_{ij}\}$
$\{D_{ij}\}$
$|S_{ij}-T_{ij}|$
0
0
0
0

Figure 11: Difference matrix between two speakers' matrices

$$D_{ij} = |S_{ij} - T_{ij}|, \quad (i < j). \tag{5}$$

In Equation 3, $\{D_{ij}\}$ are used to calculate $D_1$. In [11, 12], $\{D_{ij}\}$ were used as input features to linear regression to predict similarity between a teacher and a student. In the experiments of this paper, a part or a total of $\{D_{ij}\}$ were used as input features in SVR to predict the accent distance between $S$ and $T$. The maximum number of the features was 24,310 (=${}_{221}C_2$). $\epsilon$-SVR in LIBSVM [29] was used with the radial basis function kernel of $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$.

### 6.2. Two modes of cross-validation

Since the number of available speakers is 370, which is not very large, the following two modes of cross-validation were used in the experiments: speaker-pair-open and speaker-open.

#### *6.2.1. Speaker-pair-open cross-validation*

As our task is to predict the accent distance between two speakers, the input features to SVR have to be related to differences or distinctions in terms of pronunciation between the two speakers. In the *speaker-pair-open* mode, a training data set and a testing data set were separated so that any of a single speaker pair did not exist in both the data sets. For example, by sorting the 68,265 ($={}_{370}C_2$) speaker pairs according to the IPA-based reference distance, they can be separated based on evenness/oddness of the order. In this mode, when speaker pair A-B is found in the testing set, speaker pairs of A-$\{x\}$ ($x \neq B$) and B-$\{y\}$ ($y \neq A$) are included in the training set.

In this work, SVR was adopted as regression model. In this model, input features are mapped into a very high-dimensional feature space, where inner product between an input sample and each of all the training samples is calculated by using a kernel function. Values of inner product can be regarded as similarity scores and regression is done by using these scores as weights. When one wants to predict the accent distance between A and B, the prediction performance is supposed to be influenced by whether $\{x\}$ include a speaker who is close to B or $\{y\}$ include a speaker who is close to A in the training set.

#### *6.2.2. Speaker-open cross-validation*

It is possible to separate a training data set and a testing data set so that they do not include any of a single speaker at the same time. In the *speaker-open* mode, this strategy is adopted. When speaker pair A-B is found in the testing set, the training set includes neither of A or B. The prediction performance of SVR is supposed to be influenced by whether a speaker pair who are close enough to A and B is found in the training set.

When the number of available speakers for training SVR is $n$ and speakers A and B are testing speakers, in the speaker-pair-open mode, the prediction performance will be affected by whether a speaker close to A or B is found in the $n$ speakers. In the speaker-open mode, however, it will be affected by whether a speaker pair close to A-B is found in speaker pairs of the $n$ speakers. In other words, accent variability is estimated as $O(N)$ in the former, and in the latter, it is estimated as $O(N^2)$, where $N$ is speaker variability considered. This indicates that the latter mode requires a larger amount of training data.

#### *6.2.3. Practical interpretation of the two modes*

In the speaker-pair-open mode, either speaker of a testing speaker pair is always included in the training data. This will correspond to the following case. After training a regression model by using all the available speakers and IPA transcripts of the SAA, one wants to predict the accent distance from a new speaker to each of the SAA speakers, where IPA transcript of the new speaker is not available. The ultimate goal of this work, however, is creation of a speaker-basis accent map of World Englishes. For this goal, the accent distance between two new speakers has to be estimated adequately. The speaker-open mode is examined for this reason.

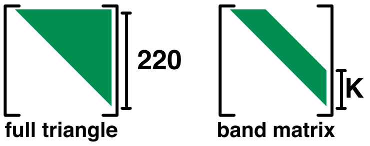

Figure 12: Paragraph-based full matrix and its band matrix

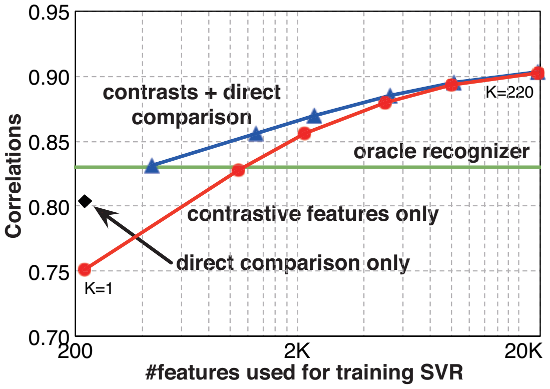

Figure 13: Distance prediction in the speaker-pair-open mode

### 6.3. Experimental conditions

Acoustic features used for training the paragraph-based UBM-HMM and adapting it were MFCC-based features: MFCC + ΔMFCC. Three states were assigned to a phoneme of the paragraph and only a single Gaussian distribution is assigned to each state of the HMM for easy calculation of BD. For pronunciation analysis, BD was calculated by using MFCC features only.

As shown in Figure 11, $D_{ij}$ is a difference between two speech contrasts of $S_{ij}$ and $T_{ij}$. Some contrasts are local but others are distant. For example, the last phoneme of the SAA paragraph is distant from the first one by 220 phonemes, which is the most distant speech contrast in $\{S_{ij}\}$ and $\{T_{ij}\}$. To investigate which contrasts contribute better, feature selection was done based on locality of speech contrasts. By using only the elements close to the diagonal of matrix $D$, i.e., band matrix, locality can be controlled. In Figure 12, $K$ is the width of the band and varies from 1 to 220. If $K$ is set to 220, it is the case where all the 24,310 elements are used as input features to SVR.

We also investigated the effectiveness of using acoustic distances obtained by direct and absolute comparison. $\{D_{ij}\}$ is results of contrastive or relative comparison, where speech contrasts are compared between two speakers. However, comparison can be done in a direct or absolute way. The paragraph-based HMM adapted to a speaker and that to another can be compared and matched directly and 221 phoneme-based BDs can be used as additional features in SVR. The features here are results of absolute comparison between two speakers. In Section 3.4, features based on absolute comparison gave us complete speaker clustering, not accent clustering. That experiment, however, was done in an unsupervised mode. SVR can run only in a supervised mode and it may show some effectiveness of using the absolute features.

### 6.4. Results and discussion

#### *6.4.1. Results in the speaker-pair-open mode*

2-fold cross-validation was done. Figure 13 shows the performances of predicting the accent distances as a function of $K$. The performance of absolute comparison only and that of our oracle baseline system are also plotted. Use of absolute fea-

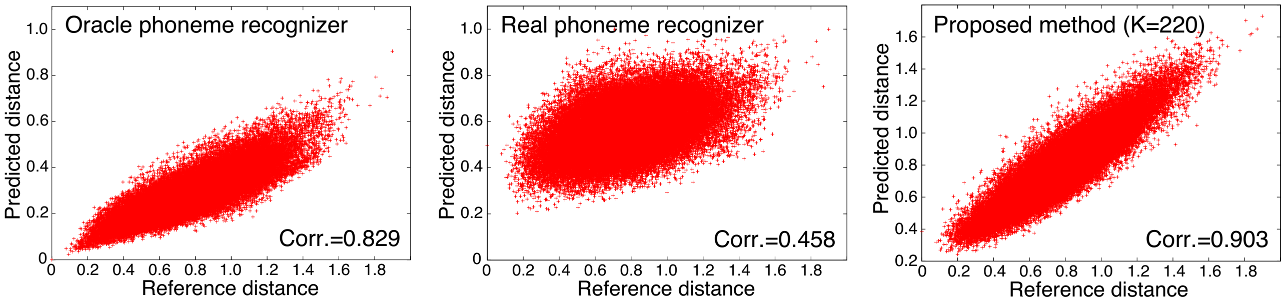

Figure 14: Correlations between the reference distances and predicted distances

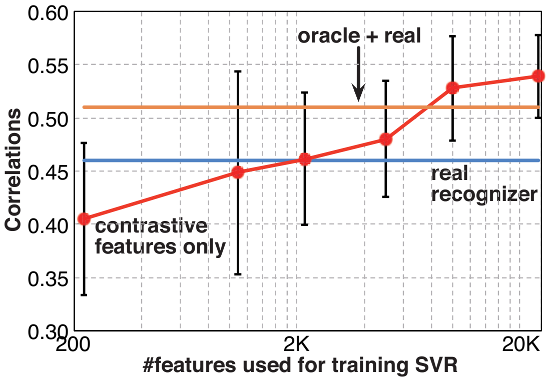

Figure 15: Distance prediction in the speaker-open mode

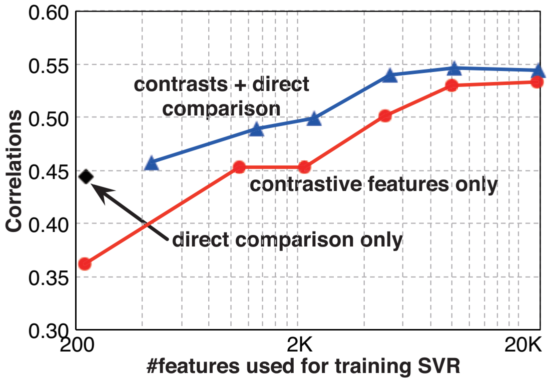

Figure 16: Effectiveness of using absolute features in the speaker-open mode

tures in SVR is unexpectedly effective and the correlation is 0.805, which is lower than the performance of our oracle system (=0.829). The number of absolute features is 221 and similar to that of contrastive features when $K$=1 (#features = 220). In this case, the correlation based on structural comparison is lower and it is 0.750. By increasing $K$, however, it improves monotonously and the maximum correlation is found when $K$=220, which is 0.903 and much higher than the performance of our oracle baseline system. Figure 13 also shows the effectiveness of combining structural features and absolute features. When $K$ is small, the effectiveness is clearly observed but when $K$ is large enough, it diminishes. We can say that very long speech contrasts still can work effectively to improve the prediction performance. Figure 14 shows correlation graphs of the three cases, a) our oracle baseline system, b) real phoneme recognizer, and c) proposed method ($K$=220).

### 6.4.2. *Results in the speaker-open mode*

5-fold cross-validation was done. Figure 15 shows the performance of structural features in predicting the accent distances where the number of features was controlled by $K$. Different from the results of the speaker-pair-open mode, the correlations are very low but the correlation of $K$=220 is sill higher than the performance of a real phoneme recognizer. This performance was obtained by automatically recognizing phonemes of spoken paragraphs of two speakers and conducting DTW between the generated phoneme transcripts. As was discussed in Section 6.2.3, at least, one of the two input speakers is included in the training data. Considering this, a new correlation was calculated, where the distance between two speakers was obtained from the oracle phoneme transcript of a speaker to the automatically generated phoneme transcript of the other. The correlation is between the new distances and the IPA-based phonetic distances and it was 0.510 (See Figure 15). Our proposed method shows a higher correlation than this.

Figure 16 shows performance improvement by adding absolute features to structural features. This result was obtained from one set of the five test sets in the 5-fold cross-validation. Here, the effectiveness is observed even when $K$ is large. When absolute features only are used, the correlation is very close to the performance of a real phoneme recognizer. We can say that structural features are very effective also in this task.

### 6.4.3. *Discussion*

The proposed method shows a striking performance of predicting the accent distance in a speaker-pair-open mode. In a speaker-open mode, however, it is not effective enough although it shows a higher performance than our second baseline system. As we discussed in Section 6.2, it is very obvious that the speaker-open mode requires a larger amount of training data. In this paper, from more than 18 hundred speakers of the SAA, available speakers were selected. In [7, 8], to increase the amount of available data, manual edition of IPA transcripts was done and filled pauses such as “ah” and “well” were manually removed from IPA transcripts. Similar operations of waveform edition are possible for spoken paragraphs to increase the amount of training data drastically in our task. Further, refinement of features for SVR such as multiple stream structuralization [11], adequate selection of kernel functions, use of multiple kernels [30], optimization of hyper-parameters of SVR, and kNN-SVR [31] will be investigated in future works.

## 7. Conclusions

This paper proposed and evaluated a method of predicting the accent distances by using the invariant pronunciation structure analysis and SVR. Reference accent distances were calculated by using narrow IPA transcripts included in the SAA corpus. Experimental results showed a very striking performance in a speaker-pair-open mode but the performance was not sufficient in a speaker-open mode. Some possible solutions are also discussed. In addition to technical and computational improvement, we're developing an i-OS application for easier collection of spoken samples using the SAA elicitation paragraph. We're not sure whether collection from 1.5 billions of speakers is an achievable goal but we're very interested in drawing a *really* global and speaker-basis map of World Englishes.

## 8. References

[1] B. Kachru *et al.*, *The handbook of World Englishes*, Wiley-Blackwell, 2009.

[2] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.

[3] H.-P. Shen, *et al.*, "Automatic pronunciation clustering using a world English archive and pronunciation structure analysis", *Proc. ASRU*, 222–227, 2013.

[4] S. Kasahara, *et al.*, "Improved and robust prediction of pronunciation distance for individual-basis clustering of World Englishes pronunciation", *Proc. ICASSP*, 2014 (to appear).

[5] T. G. Wiley, "English as a lingua franca in the age of globalization: promoting policies for co-existence and equity in a multilingual world," Keynote speech in International Conference for World Englishes, 2014.

[6] Speech Accent Archive,
http://accent.gmu.edu

[7] M. Wieling *et al.*, "A cognitively grounded measure of pronunciation distance," *PLoS ONE*, 2013 (to appear).

[8] M. Wieling *et al.*, "Automatically measuring the strength of foreign accents in English,"
http://urd.let.rug.nl/nerbonne/papers/
WielingEtAl-Accents-Validating-2013-final1.pdf

[9] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, 889–892, 2005.

[10] Y. Qiao, *et al.*, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[11] N. Minematsu, *et al.*, "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 3, 299–319, 2010.

[12] M. Suzuki, *et al.*, "Integration of multilayer regression with structure-based pronunciation assessment," *Proc. INTERSPEECH*, 586–589, 2010.

[13] The CMU pronunciation dictionary,
http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[14] N. Minematsu, *et al.*, "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, 557–560, 2004.

[15] N. Minematsu, *et al.*, "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japa-nese) database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[16] H.-P. Shen, *et al.* "Speaker-based accented English clustering using a world English archive," *Proc. SLaTE*, CD-ROM, 2013.

[17] J. Maye, *et al.*, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, 82, B101–B111, 2002.

[18] J. F. Werker, *et al.*, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, 103, 147–162, 2007.

[19] K. Okanoya, "Birdsongs and human language: common evolutionary mechanisms," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1-17-5, 1555–1556, 2008 (including Q&A after his presentation)

[20] R. Jakobson *et al.*, *The sound shape of language*, Mouton de Gruyter, 2002.

[21] W. Labov, *et al.*, *Atlas of North American English*, Mouton and Gruyter, 2005.

[22] S. Nakagawa, Y. Tohkura, and K. Shikano, *Speech, hearing and neural network*, Ohmsha, Tokyo, 1990.

[23] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, 13, 5, 930–944, 2005.

[24] D. Saito, N. Minematsu, K. Hirose, "Rotational properties of vocal tract length difference in cepstral space," *Journal of Research Institute of Signal Processing*, 15, 5, 363–374, 2011.

[25] M. D. Hauser *et al.*, "The evolution of the music faculty: a comparative perspective," *Nature neurosciences*, 6, 663–668, 2003

[26] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. ICSLP, 1669–1672, 2004.

[27] N. Minematsu, *et al.*, "Structural representation of the pronunciation and its use for classifying Japanese learners of English," Proc. SLaTE, CD-ROM, 2007.

[28] HTK Wall Street Journal Training Recipe,
http://www.keithv.com/software/htk/

[29] C.-C. Chang *et al.*, LIBSVM, a library for support vector machine, 2001.

[30] A. D. Dileep *et al.*, "Representation and feature selection using multiple kernel learning," Proc. Int. Joint Conf. Neural Networks, 717–722, 2009.

[31] W.-L. Chao *et al.*, "Facial age estimation based on label-sensitive learning and age-specific local regression," *Proc. ICASSP*, 1941–1944, 2012.